NeoBiota 59: 61–76 (2020)
doi: 10.3897/neobiota.59.36299
http://neobiota.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
NeoBiota
Advancing research on alien species and biological invasions

Origin of climatic data can determine the transferability of species distribution models

Arunava Datta[1,2,3], Oliver Schweiger[1], Ingolf Kühn[1,4,5]

1 *Department of Community Ecology, Helmholtz Centre for Environmental Research–UFZ, Theodor-Lieser-Straße 4, D-06120 Halle, Germany* **2** *Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Matieland, South Africa* **3** *South African National Biodiversity Institute, Kirstenbosch National Botanical Gardens, Claremont, South Africa* **4** *Institute of Biology/Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, D-06108 Halle, Germany* **5** *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany*

Corresponding author: Arunava Datta (arunava.datta@ufz.de)

Academic editor: M. Rejmanek | Received 19 May 2019 | Accepted 31 May 2020 | Published 28 July 2020

Citation: Datta A, Schweiger O, Kühn I (2020) Origin of climatic data can determine the transferability of species distribution models. NeoBiota 59: 61–76. https://doi.org/10.3897/neobiota.59.36299

Abstract

Methodological research on species distribution modelling (SDM) has so far largely focused on the choice of appropriate modelling algorithms and variable selection approaches, but the consequences of choosing amongst different sources of environmental data has scarcely been investigated. Bioclimatic variables are commonly used as predictors in SDMs. Currently, several online databases offer the same sets of bioclimatic variables, but they differ in underlying source of raw data and method of data processing (extrapolation and downscaling). In this paper, we asked whether predictive performance and spatial transferability of SDMs are affected by the choice of two different bioclimatic databases viz. WorldClim 2 and Chelsa 1.2. We used presence-absence data of the invasive plant *Ageratina adenophora* from the Western Himalaya for training SDMs and a set of independently-collected presence-only datasets from the Central and Eastern Himalaya to evaluate the transferability of the SDMs beyond the training range. We found that the performance of SDMs was, to a large degree, affected by the choice of the climatic dataset. Models calibrated on Chelsa 1.2 outperformed WorldClim 2 in terms of internal evaluation on the calibration dataset. However, when the model was transferred beyond the calibration range to the Central and Eastern Himalaya, models based on WorldClim 2 performed substantially better. We recommend that, in addition to the choice of predictor variables, the choice of predictor datasets with these variables should not be based merely on subjective decision whenever several options are available. Instead, such decisions should be based on robust evaluation of the most appropriate dataset for a given geographic region and species being modelled. Moreover, decisions could also depend on the objective of the study, i.e. projecting within the calibration range or beyond. Therefore, a quantitative evaluation of predictor datasets from alternative sources should be routinely performed as an integral part of the modelling procedure.

Copyright *Arunava Datta et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Keywords
Ageratina adenophora, climatic database, invasive species, model transfer, species distribution modelling

Introduction

Correlative species distribution models (SDMs, also referred to as ecological niche models or habitat suitability models) are used to estimate the potential geographic distribution of species by modelling the relationship between known occurrences of a species with its environmental conditions (Guisan and Zimmermann 2000; Pearson and Dawson 2003; Elith and Leathwick 2009). These models directly relate the occurrence of a species to its realised multi-dimensional niche (Hutchinson 1957; Pearson and Dawson 2003) in the environmental space (Soberón and Nakamura 2009; Peterson et al. 2011) that is provided by the chosen predictor variables. Climatic conditions are crucial in determining the large-scale distribution patterns of organisms (Woodward 1987; Woodward et al. 2004) and are hence widely used for modelling species distributions (Pearson and Dawson 2003).

SDMs are frequently applied in invasion biology, conservation biology, evolutionary biology and agriculture due to their versatility (Elith and Leathwick 2009; Peterson et al. 2011). SDMs of invasive species are often used to make temporal and spatial predictions of climatically-suitable regions that could potentially be invaded (Thuiller et al. 2005; Ervin and Holly 2011; Jaryan et al. 2013) and thus aid in early detection, control and eradication of the invasive species (Thuiller et al. 2005; Peterson et al. 2011). The distribution of invasive plants will most likely change due to climate change and therefore future projections of invasion from SDMs will further help in taking long-term management decisions (Thuiller et al. 2005; Peterson et al. 2011).

To avoid misleading recommendations for such management decisions, SDMs and the resulting predictions or future projections of suitable environmental conditions and corresponding invasion risks need to be highly reliable. Much of past research has focused on the development of modelling algorithms and model (i.e. variable) selection to increase the performance of SDMs (Guisan and Zimmermann 2000; Elith and Leathwick 2009). Ample studies are available on different methodological aspects, such as the choice of different modelling algorithms, sample size, sample density, variable selection and spatial resolution of environmental layers on model accuracy and transferability (Randin et al. 2006; Peterson and Nakazawa 2008; Heikkinen et al. 2012; Wenger and Olden 2012).

Model transferability, either in space or time (Randin et al. 2006; Elith and Leathwick 2009), is of particular importance for invasive species to reliably assess their response to climate change or to predict their invasive potential in novel areas and for corresponding management decisions (Clark et al. 2001; Yates et al. 2018). Therefore, it is essential to assess the predictive accuracy of an SDM, not only within the region in which it was fitted (i.e. internal validation within the calibration range), but also in a geographic region different from the calibration range (i.e. external validation on an

independent dataset) (Heikkinen et al. 2012; Wenger and Olden 2012; Fernández and Hamilton 2015). The model transfer may often involve extrapolation if the ranges of the predictors are beyond the calibration range of the model. Model transferability is a particularly challenging issue in species distribution modelling (Araújo and Guisan 2006; Elith and Leathwick 2009; Peterson et al. 2011; Wenger and Olden 2012). A recent review on challenges in transferability of ecological models has flagged many pertinent issues, such as the choice of response variables, sampling bias, choice of modelling algorithm and non-stationarity etc. (Yates et al. 2018).

It has also been shown that the choice of predictor variables can impact model accuracy and transferability (Bobrowski et al. 2017; Karger et al. 2017; Petitpierre et al. 2017), but studies, focusing exclusively on the consequences of choosing different sources providing the same set of predictor variables, are very scarce (Peterson et al. 2011). Consequently, researchers often rely on their subjective decisions for choosing one source of predictor datasets over others, even if the same set of (potential predictor) variables are available from different sources.

SDMs have increasingly benefitted from the availability of climatic predictors at very high resolutions in the form of rasterised GIS layers available from different sources (Soberón and Nakamura 2009; Peterson et al. 2011). Despite offering the same variables, such different climatic databases could differ in their actual values since they rely on different source data and use different interpolation or downscaling algorithms (Bobrowski and Schickhoff 2017; Karger et al. 2017). Such differences could be particularly relevant in regions of high orographic heterogeneity, which have been shown to be highly sensitive to prediction errors for multiple plant species (Hanspach et al. 2011).

The most widely-used variables for SDMs are the set of 19 bioclimatic variables (Peterson and Nakazawa 2008; O'Donnell and Ignizio 2012) that do not only include annual averages, but also climatic extremes limiting the physiological performance of biological organisms (O'Donnell and Ignizio 2012). Currently, several databases offer free access to these bioclimatic variables. WorldClim was one of the first and most frequently used high resolution (30 arc seconds) global bioclimatic dataset derived from ground weather stations across the globe and interpolated by using latitude, longitude and elevation as independent variables (Hijmans et al. 2005). In the recent version of WorldClim (Version 2; Fick and Hijmans 2017), hereafter referred to as WorldClim 2, satellite-derived covariates, such as land surface temperature and cloud cover, have also been used in the interpolation process to improve the data quality in areas where ground observations are scarce. Chelsa (Version 1.2; Karger et al. 2017), hereafter referred to as Chelsa 1.2, is another bioclimatic database that accounts for orographic patterns of precipitation in mountainous terrains, i.e. it accounts for factors, such as aspect and valley exposition by including wind effects (see Karger et al. 2017). Therefore, it can be assumed that, due to the methodological differences in generating the raster layers, these databases are not equivalent and hence their use in SDMs could result in differences in predictive accuracy and, moreover, in transferability.

In this paper, we asked, whether models calibrated on Chelsa 1.2 and WorldClim 2, respectively, differ in terms of internal and external predictive performance. To this end, we used the invasive plant species *Ageratina adenophora* (Spreng.) R.M.King &

H.Rob. in the Himalaya as our study system. Using presence-absence data of *A. adenophora* from the Western Himalaya as the response, we calibrated generalised linear models on Chelsa1.2 and WorldClim2 data. Transferability of models calibrated on these two datasets was evaluated using an independent set of presence-only data from Central and Eastern parts of the Himalaya.

Methods

Target species

Ageratina adenophora (Crofton weed, Asteraceae) is a plant species native to Mexico and invasive (or even noxious) in more than 30 countries in subtropical regions across the globe (Auld and Martin 1975; Qiang 1998; Tian et al. 2007; Muniappan et al. 2009; Poudel et al. 2019). It is a multi-stemmed, perennial herb or undershrub that grows up to 2 metres and flowers profusely in spring (Tripathi et al. 2012). It was introduced as an ornamental plant to England in the 19th century (Auld and Martin 1975) and was later introduced in different parts of the world (Muniappan et al. 2009), such as the Himalaya (Dehradun, India) in the early 20th century (Datta et al. 2017). In South Asia, it has expanded its distribution almost throughout the subtropical and sub-temperate belts of the Himalaya, ranging from Arunachal Pradesh in the east to Himachal Pradesh in the west (Raizada 1976; Tripathi et al. 2012) and also flourishes in mountains of peninsular India (Muniappan and Viraktamath 1993; Muniappan et al. 2009).

Study area and distribution survey

Our study was carried out in a region of the Western Himalaya (Singh and Singh 1987) between 29.96N and 32.55N latitudes and 75.77E and 78.43E longitudes. Our study area covered five provinces in north-western India and stretched from Dhauladhar range (Himachal Pradesh province) in the west to the mountains of Gharwal region (Uttrakhand province) in the east. We also covered a considerable part of low-lying foothills of the Himalaya (Siwalik range).

We haphazardly surveyed 389 locations and recorded the presence or absence of *A. adenophora* in the subtropical and temperate zones of the Western Himalaya between 300 m to 3000 m elevation (Fig.1). We targeted this elevational belt based on prior knowledge about the distribution of the plant from previous reconnaissance surveys and existing literature on its distribution (Datta et al. 2017). The surveys were conducted in the vegetation periods of 2014 and 2015. Most of the surveys were carried out along road- and riversides as these are conduits for dispersal of propagules and are also initial establishment sites of *A. adenophora* (Z. Lu and Ma 2006; Wang et al. 2011). However, many high elevational areas beyond 2500 m were not accessible by road and, hence, we used trekking trails for surveying such remote locations. Alpine

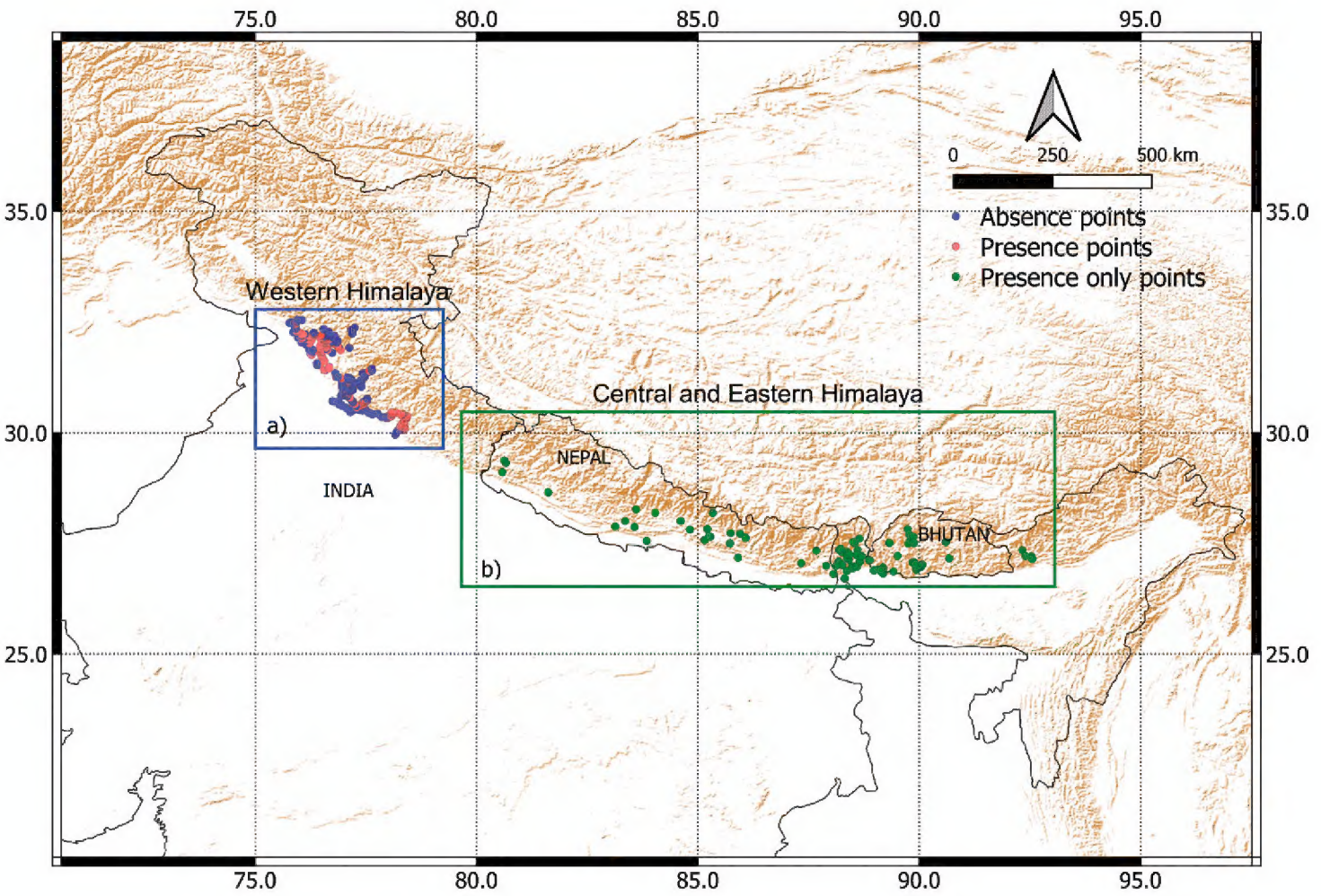

Figure 1. Survey locations of *Ageratina adenophora*. The region marked by the blue rectangle **a** shows the survey area in the Western Himalaya from which 192 presences (red circles) and 197 genuine absences (blue circles) were used to train the model. The region marked by the green rectangle **b** shows the Central and Eastern Himalaya from where an additional set of 85 presence only locations (green circles) were obtained for evaluating the transferability of the species distribution models trained in the Western Himalaya.The relief map of the region is depicted in brown. The relief map was made with layer obtained from Natural Earth and the international borders were digitized from political map of India (9th edition) published by survey of Inida..

and subalpine regions (> 3500 m) were not surveyed since the plant is known to be entirely absent from these regions due to extremely low temperatures (Datta et al. 2017). At the scale of the used climatic variables (30 arc seconds or 1 km^2), microclimatic variations due to trails, roads and water conduits are not a limiting factor for the distribution of the species. Hence, this potential bias in data acquisition should not influence the model outcome and general conclusions. To assess model transferability, we used an independent set of presence-only records (N = 85) that were collected by experts (see acknowledgements) from Central and Eastern Himalaya (Fig. 1).

Climatic data and variable selection

Due to collinearity amongst the 19 bioclimatic variables, we used a cluster analysis to select variables seperately for WorldClim 2 and Chelsa 1.2 (Dormann et al. 2013). All the 19 bioclimatic variables were scaled to zero and unit standard deviation prior to

cluster analysis. The dendrogram was constructed, based on Spearman's rank correlation (ρ) using UPGMA (unweighted pair-group method with arithmetic averages) agglomeration method. A threshold value of of ρ = |0.7| (Dormann et al. 2013) was used to prune the dendrogram and select variables that were not highly collinear. This procedure resulted in five clusters for WorldClim1.2 and seven clusters for Chelsa 2 (see Suppl. material 1). Selection of a variable within a cluster was primarily based on its ecological relevance to the study species. For example, the plant is known to be limited by low temperatures in higher elevations (Datta et al. 2017), therefore the minimum temperature of the coldest month was selected (bio 6). Similarly, germination of seeds is known to be limited by moisture in the lower elevations, hence precipitation of the driest month (bio 14) was preferred over other variables (Datta et al. 2017). In order to make the models based on WorldClim 2 and Chelsa 1.2 comparable, we ensured that the set of selected variables was common. However, due to inherent differences in the correlation sructure, the variable selection procedure yielded slightly different sets of variables for the two datasets. Finally, five variables were selected for WorldClim 2, while two additional variables were selected in the case of Chelsa 1.2 (see Table 1).

In addition to the two models based on WorldClim 2 and Chelsa 1.2 data, we calibrated a third model based on Chelsa 1.2 data, but using the same set of five variables that were selected specifically for WorldClim 2 (Table 1). This allowed us to make direct and unbiased comparison between the predictive performance of WorldClim 2 and Chelsa 1.2 and to assess whether our conclusions are potentially confounded by differences in model performance caused by the initial variable selection procedure.

Modelling procedure

We used a multi-model inference approach (Burnham and Anderson 2002) to arrive at the final model to be used for prediction (Grueber et al. 2011; Symonds and Moussalli 2011; Burnham 2015). The following steps were taken: (1) We fitted generalised linear models with binomial error distribution and a logit link function to the presence-absence data of *A. adenophora* in the Western Himalaya using previously selected climatic variables (Table 1). All predictor variables were scaled to zero mean and unit standard deviation. (2) We then obtained models with all possible variable combinations using the "dredge" function in the "MuMIn" package (Barton 2015) of R (R Core Team 2017). (3) A subset of best models that differed by 2 or less in AIC from the best model was considered for a model averaging process (hereafter referred to as "best subset") (Grueber et al. 2011). (4) We then averaged model coefficients, weighted by the corresponding Akaike weights across all models in the best subset. We used the default "full average" method for calculating the averaged coefficients (if a variable is absent from one of the component models, a parameter estimate of "zero" is substituted in the averaging process (Symonds and Moussalli 2011). This method results in shrinkage of parameter estimates for those variables which are less important (Grueber et al. 2011) and has been suggested when prediction from an averaged model is intended (Symonds and Moussalli 2011).

Table 1. Variable selection for Chelsa 1.2 and WorldClim 2 datasets using UPGMA cluster analysis to reduce collinearity amongst the variables. Highly correlated variables were removed from each dataset (using threshold of Spearman's **ρ** = 0.7, see text for details). The selected variables from Chesla 1.2 and WorldClim 2 are represented by tick mark (✓) against the respective variable.

Climatic variable	Abbreviation	WorldClim2	Chelsa1.2
Isothermality	bio3	✓	✓
Temperature Seasonality	bio4		✓
Min Temperature of Coldest Month	bio6	✓	✓
Temperature Annual Range	bio7		✓
Annual Precipitation	bio12	✓	✓
Precipitation of Driest Month	bio14	✓	✓
Precipitation Seasonality	bio15	✓	✓

Model evaluation

To obtain binary predictions (i.e. presence or absence output) from continuous probability values, a threshold was selected by maximising the true skill statistic (TSS), which accounts for both omission and commission errors and is known to be independent of prevalence (Allouche et al. 2006). The value of TSS can range from -1 to +1. A value close to +1 indicates good agreement, while a value close to 0 indicates that the model does not perform better than a random model (Allouche et al. 2006). A value close to -1 suggests that a completely inverse model would be better. AUC is a common traditionally-used metric for evaluating the performance of SDMs; however, its efficiency has been questioned (Jiménez-Valverde et al. 2008) and, therefore, we do not report AUC values.

To assess the transferability (i.e. predictive performance of the model beyond the calibration area in the Western Himalaya), we used the independent set of presence-only data from the Central and Eastern Himalaya (Nepal, Sikkim, Darjeeling and Bhutan; see acknowledgements for contributors). Since we did not have true absence data from these regions, we could not use ordinary model evaluation metrics such as TSS. Therefore, we used the Boyce Index for assessing transferability (Boyce et al. 2002; Hirzel et al. 2006). The Boyce Index compares the ratio of predicted frequency and expected frequency of evaluation points across the prediction gradient using a moving window approach (Hirzel et al. 2006; Petitpierre et al. 2012). It is a threshold-independent metric ranging between -1 and +1. Positive values close to 1 indicate very good agreement of observed presences with the model prediction, while values very close to zero indicate that the predictions are not better than random. Negative values of the Boyce Index show that the model is worse than a random model and makes predictions in areas that are not suitable for the species (Hirzel et al. 2006). For this purpose, the region of evaluation was defined by drawing a convex hull around the presence-only evaluation points. The convex hull (polygon) was used to crop the prediction layer (raster) from the model. Subsequently, the predicted occurrence probabilities were used as a measure of "habitat suitability" (x-axis) and were correlated

(Spearman's correlation) with the "predicted to expected ratio" (y-axis) calculated from the presence-only evaluation points across the prediction gradient using the moving window approach (Hirzel et al. 2006).

The Boyce Index was calculated using the "ecospat.boyce" function of the "ecospat" package (Cola et al. 2017). The Boyce Index was also calculated for internal evaluation (i.e. training range) to facilitate direct comparison between Western and Central and Eastern Himalaya using presence only data.

Further, SDMs were projected to a much larger geographic area (entire South Asia) compared to the training area to allow for a general qualitative assessment (i.e. visual agreement), based on a priori knowledge about the distribution of *A. adenophora* from existing literature. R codes for the entire analysis can be found in Suppl. material 3.

Results

Here, we report the predictive performance of the three averaged models using the multimodel inference approach. The first model ("WorldClim data – WorldClim variable selection") had two component models (i.e. best subset of models that differed by 2 or less in AIC), the second model ("Chelsa data – Chelsa variable selection") had six component models, while the third model ("Chelsa data – WorldClim variable selection") had four component models. The average value of the coefficients for the bioclimatic variables also differed between the models (Suppl. material 2).

Internal evaluation of the models based on TSS, using presence-absence data, showed that Chelsa performed marginally better than WorldClim (Table 2). The "Chelsa data – Chelsa variable selection" had the highest TSS value amongst all models, while the "Chelsa data – WorldClim variable selection" performed similar to "WorldClim data – WorldClim variable selection" (Table 2). In contrast, internal evaluation using the Boyce Index (based on presence-only data) revealed that the performance of

Table 2. Model evaluation metrics for different models using Chelsa 1.2 and WorldClim 2 datasets. Database refers to the climatic database used for modelling (calibration). Variable selection refers to the specific set of variables selected using cluster analysis for Chelsa 1.2 and WorldClim 2 datasets (see Table 1 and method section for further details). Sensitivity is the rate of true positives while specificity is the rate of true negatives. Boyce internal refers to the Boyce Index calculated for the training area and Boyce external refers to the Boyce Index calculated for Central and Eastern Himalaya where the model was transferred to. Chelsa 1.2 and WorldClim 2 are written as Chelsa and WorldClim in the table.

		Internal evaluation							External evaluation
Modelling database	Variable selection	Thr	PCC	Sen	Spe	TSS	MSE	Boyce index	Boyce index
WorldClim	WorldClim	0.69	0.76	0.6	0.92	0.52	0.24	0.61	0.64
Chelsa	Chelsa	0.46	0.81	0.76	0.86	0.62	0.19	0.59	-0.14
Chelsa	WorldClim	0.54	0.75	0.73	0.77	0.51	0.25	0.91	0.37

Thr: Threshold to translate continuous occurrence probabilities into presence/absence data; Sen: Sensitivity; Spe: Specificity; PCC: Percent correctly classified; TSS: True skill statistic; MSE: Mean square error.

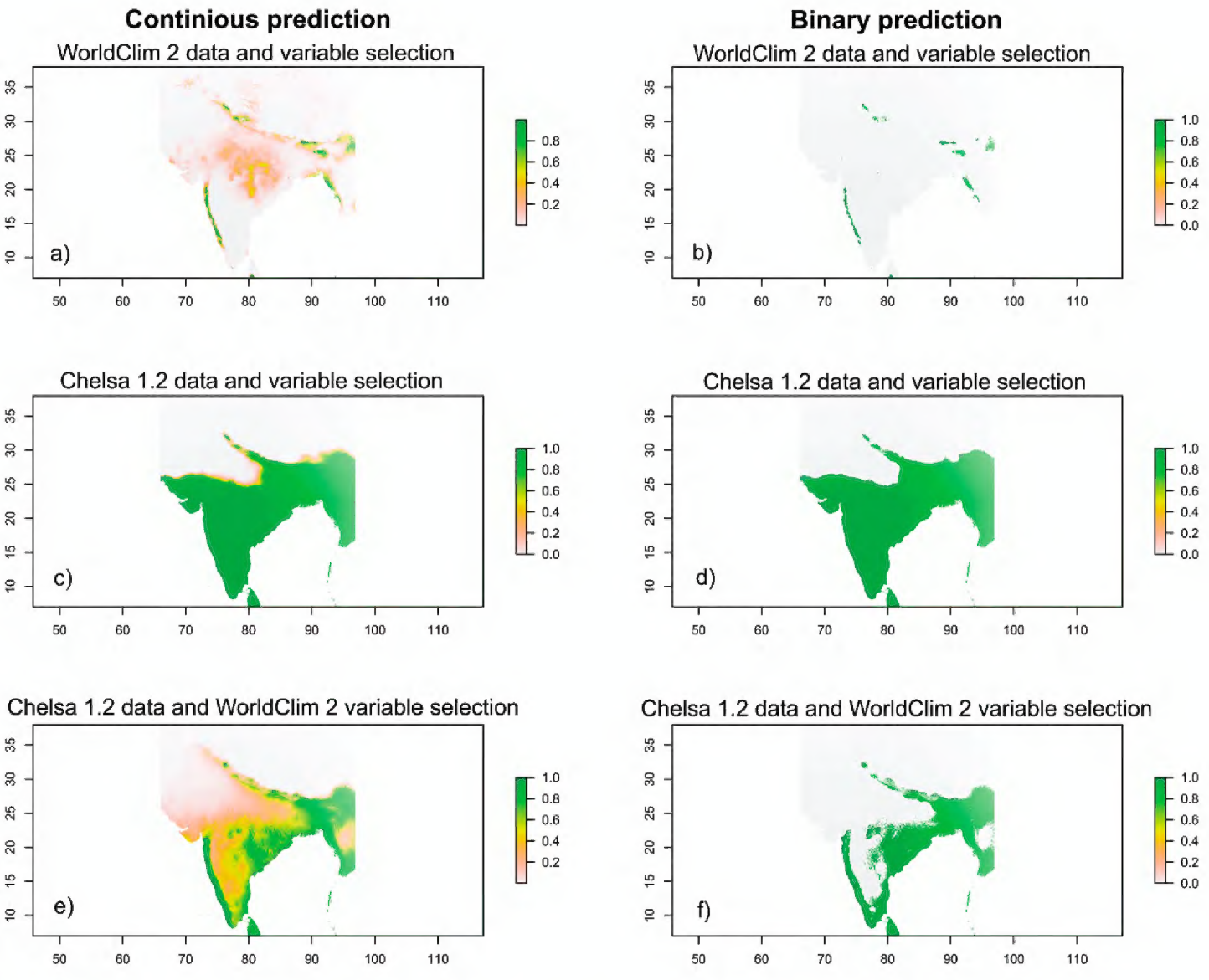

Figure 2. Model projection in South Asia showing the continuous probabilities (left) and binarised prediction (right) from the models. Panel **a** and **b**: WorldClim 2 data and variables selected for WorldClim 2; panel **c** and **d**: Chelsa 1.2 data and variables selected for Chelsa 1.2; panel **e** and **f**: WorldClim 2 data but variables selected for Chelsa 1.2.

the Chelsa models was marginally lower than the WorldClim models, while "Chelsa data – WorldClim variable selection" had the highest Boyce Index (Table 2).

In contrast to internal model evaluation, transferability of the model beyond the calibration range in the Central and Eastern Himalaya was entirely based on the Boyce Index because we had only presence data from these regions. The Boyce Index was highest for the "WorldClim data – WorldClim variable selection" and was slightly negative for "Chelsa data – Chelsa variable selection". Negative value of Boyce's Index indicated that the model predicted high probability of occurrence even for regions that were almost unsuitable for the species.

The visual inspection of the prediction maps also showed that the "Chelsa data – Chelsa variable selection" model produced extremely unrealistic over-predictions (Fig. 2c). For instance, the model showed most parts of South Asia to be highly suitable for *A. adenophora*, including warm tropical regions of peninsular India. However, in reality, the species is known to be restricted to moist subtropical and temperate regions found at higher elevations (Muniappan and Viraktamath 1993).

To identify whether this over-prediction was simply caused by the selection of variables based on the Chelsa dataset, we also assessed the performance of the "Chelsa data – WorldClim variable selection" model. This increased model performance, measured with the Boyce Index, but stayed considerably below that of the "WorldClim data – WorldClim variable selection" model (Table 2). Further, transferability was slightly improved, although many potentially unsuitable regions in central and southern India were still being predicted as climatically suitable for *A. adenophora* (Fig. 2d).

Discussion

Using two openly-available bioclimatic datasets, we found that the choice of the climatic dataset had a substantial effect on transferability of SDMs in mountainous regions such as the Himalaya. It is interesting to note that, although the same set of five variables was used in the multimodel inference approach for "WorldClim data – WorldClim variable selection" and "Chelsa data – WorldClim variable selection" models, the number of component models in the "best subset" for "Chelsa data – WorldClim variable selection" was twice the number of models in "WorldClim data – WorldClim variable selection". The contribution of the variables in these two models also differed considerably. For example, in the "WorldClim data – WorldClim variables" model, bio15 was the most important variable, but in the case of "Chelsa data – WorldClim variables", bio12 was the most important variable. This suggests that the difference in predictive power between the two databases is most likely due to the underlying differences in the variables and not due to the modelling approach used by us.

We initially expected that the Chelsea 1.2 dataset would perform very well in mountainous areas because it corrects for orographic patterns of precipitation. Earlier studies, based in the Himalaya and the Swiss Alps, showed that the performance of Chelsa was superior to WorldClim. For example, it has been reported that Chelsa 1 outperformed WorldClim 1.4 in predicting the distribution of tree line forming Himalayan birch in the Himalaya (Bobrowski and Schickhoff 2017). Karger et al. (2017) also found a marginally superior performance of Chelsa 1 over WorldClim 1.4 in predicting the distribution of 67 plant species from Switzerland. However, unlike our study, none of the previous studies verified the transferability of the models in space using an independent occurrence dataset from a different geographic region.

Our study yielded contrasting results, especially in terms of reliability when models are transferred to other regions. This difference could partly be due to the following reasons: i) earlier studies used older versions of the two climatic databases. WorldClim has considerably updated their data in the latest version (WorldClim 2) by incorporating remotely-sensed variables, such as land surface temperature and cloud cover. This update might have significantly improved the quality of the data in contrast to previous versions. ii) since Chelsa 1.2 has made several corrections to account for orographic patterns, especially in precipitation (Karger et al. 2017), these corrections might have changed the spatial pattern of the correlation structure amongst the variables at a local scale (Mesgaran et al. 2016). Therefore, the transferability of the model might be com-

promised when the models are projected to a new region characterised by a different correlation structure amongst the variables.

It is worth noting that the values of TSS were not very high for any of the models, indicating that climatic variables alone are not sufficient in explaining the distribution pattern of *A. adenophora*. For example, empirical studies have shown that the species has a narrow pH range from slightly acidic to neutral soils (pH 5 to 7) and cannot tolerate highly saline conditions (Lu et al. 2006). Moreover, biotic interactions and dispersal limitations are also crucial in determining plant distributions (Soberón and Nakamura 2009; Peterson et al. 2011). Therefore, including such variables could probably help in improving the general model performance and transferability for this species.

Although we found WorldClim 2 to perform better in terms of model transferability, it is premature to give generalised recommendations for preferring one dataset over the other, based on this case study alone. The species being studied and the geographic area of the study may be equally important (Hanspach et al. 2011). Providing a general overview, on how pertinent the problem is or under which conditions it applies for which type of species is beyond the scope of this study. We rather want to highlight the potential problem. We therefore recommend that the evaluation of climatic datasets should be performed routinely as an integral part of a modelling exercise and the database with best predictive performance should be chosen. For application of SDMs within the training and calibration region, internal validation is reliable, although performing an out-of-area cross-validation procedure is preferable when sample size is sufficient (Wenger and Olden 2012). However, if model transfer to a different geographic region is desired, validation against an independent occurrence dataset is highly recommended for choosing the most appropriate source of environmental data for the given study system. Therefore, a quantitative evaluation of predictor datasets from alternative sources should be routinely performed as an integral part of the modelling procedure.

Data availability

The occurrence data can be found here: https://zenodo.org/record/3875679#.Xtg6IzozZRZ [https://doi.org/10.5281/zenodo.3875679]

Acknowledgements

We carried out this work with financial support from German Academic Exchange Service (DAAD) and institutional support from CSIR-Institute of Himalayan Bioresource Technology, Palampur and Helmholtz Centre for Environmental Research-UFZ. For contributing occurrence data, we would like to specifically thank Dr. Rajendra Yonzone from Darjeeling (India), Choki Gyeltshen from Bhutan, Bharat Pradhan from Sikkim (India), Dr. Dinesh Thakur from Jammu (India), Om Prakash from Palampur, and Dr. Bharat Shrestha from Nepal. Finally we would like to thank Dr. R.D Singh (deceased) for his motivation to carry out the field work in Himalayas.

References

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43: 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. Journal of Biogeography 33: 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Auld B, Martin PM (1975) The autecology of *Eupatorium adenophorum* Spreng. in Australia. Weed Research 15: 27–31. https://doi.org/10.1111/j.1365-3180.1975.tb01092.x

Barton K (2015) MuMIn: Multi-model inference. R package version 1.9.13. Version 1: 18.

Bobrowski M, Gerlitz L, Schickhoff U (2017) Modelling the potential distribution of *Betula utilis* in the Himalaya. Global Ecology and Conservation 11: 69–83. https://doi.org/10.1016/j.gecco.2017.04.003

Bobrowski M, Schickhoff U (2017) Why input matters: Selection of climate data sets for modelling the potential distribution of a treeline species in the Himalayan region. Ecological Modelling 359: 92–102. https://doi.org/10.1016/j.ecolmodel.2017.05.021

Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA (2002) Evaluating resource selection functions. Ecological Modelling 157: 281–300. https://doi.org/10.1016/S0304-3800(02)00200-4

Burnham KP (2015) Multimodel Inference: Understanding AIC relative variable importance values.

Burnham KP, Anderson DR (2002) Springer Model Selection and Multimodel Inference: a Practical Information-theoretic Approach. 488 pp.

Clark JS, Carpenter SR, Barber M, Collins S, Dobson A, Foley JA, Lodge DM, Pascual M, Pielke Jr R, Pizer W, Pringle C, Reid WV, Rose KA, Sala O, Schlesinger WH, Wall DH, Wear D (2001) Ecological Forecasts: An Emerging Imperative. Science 293: 657–661. https://doi.org/10.1126/science.293.5530.657

Cola V Di, Broennimann O, Petitpierre B, Breiner FT, Amen MD, Randin C, Engler R, Pottier J, Pio D, Dubuis A, Pellissier L, Mateo G, Hordijk W, Salamin N, Guisan A (2017) ecospat: an R package to support spatial analyses and modeling of species niches and distributions. Ecography 40(6): 774–787. https://doi.org/10.1111/ecog.02671

Datta A, Kühn I, Ahmad M, Michalski S, Auge H (2017) Processes affecting altitudinal distribution of invasive *Ageratina adenophora* in western Himalaya: The role of local adaptation and the importance of different life-cycle stages. PloS ONE 12(11): e0187708. https://doi.org/10.1371/journal.pone.0187708

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, Mcclean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36: 027–046. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Elith J, Leathwick JR (2009) Species Distribution Models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40: 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ervin GN, Holly DC (2011) Examining local transferability of redictive species distribution models for invasive plants: an example with Cogongrass (*Imperata cylindrica*). Invasive Plant Science and Management 4: 390–401. https://doi.org/10.1614/IPSM-D-10-00077.1

Fernández M, Hamilton H (2015) Ecological niche transferability using invasive species as a case study. PLoS ONE 10(3): e0119891. https://doi.org/10.1371/journal.pone.0119891

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37: 4302–4315. https://doi.org/10.1002/joc.5086

Grueber CE, Nakagawa S, Laws RJ, Jamieson IG (2011) Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology 24: 699–711. https://doi.org/10.1111/j.1420-9101.2010.02210.x

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135: 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Hanspach J, Kühn I, Schweiger O, Pompe S, Klotz S (2011) Geographical patterns in prediction errors of species distribution models. Global Ecology and Biogeography 20: 779–788. https://doi.org/10.1111/j.1466-8238.2011.00649.x

Heikkinen RK, Marmion M, Luoto M (2012) Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35: 276–288. https://doi.org/10.1111/j.1600-0587.2011.06999.x

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965–1978. https://doi.org/10.1002/joc.1276

Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling 199: 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017

Hutchinson GE (1957) Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology 22: 415–427. https://doi.org/10.1101/SQB.1957.022.01.039

Jaryan V, Datta A, Uniyal SK, Kumar A, Gupta RC, Singh RD (2013) Modelling potential distribution of *Sapium sebiferum* – an invasive tree species in western Himalaya. Current Science 105: 1282–1287.

Jiménez-Valverde A, Lobo JM, Hortal J (2008) Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14: 885–890. https://doi.org/10.1111/j.1472-4642.2008.00496.x

Karger DN, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza RW, Zimmermann NE, Linder HP, Kessler M (2017) Climatologies at high resolution for the earth's land surface areas. Scientific Data 4: 170122. https://doi.org/10.1038/sdata.2017.122

Lu P, Sang W, Ma K (2006) Effects of environmental factors on germination and emergence of Crofton weed (*Eupatorium adenophorum*). Weed Science 54: 452–457. https://doi.org/10.1614/WS-05-174R1.1

Lu Z, Ma K (2006) Spread of the exotic croftonweed (*Eupatorium adenophorum*) across southwest China along roads and streams. Weed Science 54: 1068–1072. https://doi.org/10.1614/WS-06-040R1.1

Mesgaran MB, Lewis MA, Ades PK, Donohue K, Ohadi S, Li C (2016) Hybridization can facilitate species invasions , even without enhancing local adaptation. Proceedings of the National Academy of Sciences 113: 10210–10214. https://doi.org/10.1073/pnas.1605626113

Muniappan R, Raman A, Reddy GVP (2009) *Ageratina adenophora* (Sprengel) King and Robinson (Asteraceae). Biological Control of Tropical Weeds using Arthropods. Cambridge University Press, 63–73. https://doi.org/10.1017/CBO9780511576348.004

Muniappan R, Viraktamath CA (1993) Invasive alien weeds in the Western Ghats. Current Science 64: 555–558.

O'Donnell MS, Ignizio DA (2012) Bioclimatic predictors for supporting ecological applications in the conterminous United States. U.S Geological Survey Data Series 691. https://doi.org/10.3133/ds691

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12: 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Peterson AT, Nakazawa Y (2008) Environmental datasets matter in ecological niche modelling: An example with Solenopsis invicta and Solenopsis richteri. Global Ecology and Biogeography 17: 135–144. https://doi.org/10.1111/j.1466-8238.2007.00347.x

Peterson AT, Soberón J, Pearson RG, Anderson RP, Martínez-Meye E, Nakamura M, Araújo MB (2011) Ecological Niches and Geographic Distributions (MPB-49). Princeton University Press. https://doi.org/10.23943/princeton/9780691136868.001.0001

Petitpierre B, Broennimann O, Kueffer C, Daehler C, Guisan A (2017) Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. Global Ecology and Biogeography 26: 275–287. https://doi.org/10.1111/geb.12530

Petitpierre B, Kueffer C, Broennimann O, Randin C, Daehler C, Guisan A (2012) Climatic niche shifts are rare among terrestrial plant invaders. Science (New York, N.Y.) 335: 1344–1348. https://doi.org/10.1126/science.1215933

Poudel AS, Jha PK, Shrestha BB, Muniappan R (2019) Biology and management of the invasive weed *Ageratina adenophora* (Asteraceae): current state of knowledge and future research needs. Weed Research 59(2): 79–92. https://doi.org/10.1111/wre.12351

Qiang S (1998) The history and status of the study on croftonweed (*Eupatorium adenophorum* Spreng.) A worst worldwide weed. Journal of Wuhan Botanical Research 16: 366–372.

Raizada MB (1976) Supplement to Duthie's flora of the ppper Gangetic plain and of adjacent Siwalik and sub- Himalayan tracts. Bishen Singh Mahendra Pal Singh, Dehradun, 358 pp.

Randin CF, Dirnböck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006) Are niche-based species distribution models transferable in space? Journal of Biogeography 33: 1689–1703. https://doi.org/10.1111/j.1365-2699.2006.01466.x

Singh JS, Singh SP (1987) Forest vegetation of the Himalaya. The Botanical Review 53: 80–192. https://doi.org/10.1007/BF02858183

Soberón J, Nakamura M (2009) Niches and distributional areas: Concepts, methods, and assumptions. Proceedings of the National Academy of Sciences 106: 19644–19650. https://doi.org/10.1073/pnas.0901637106

Symonds MRE, Moussalli A (2011) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology 65: 13–21. https://doi.org/10.1007/s00265-010-1037-6

Thuiller W, Richardson DM, Pyšek P, Midgley GF, Hughes GO, Rouget M (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Global Change Biology 11: 2234–2250. https://doi.org/10.1111/j.1365-2486.2005.001018.x

Tian Y, Feng Y, Liu C (2007) Addition of activated charcoal to soil after clearing *Ageratina adenophora* stimulates growth of forbs and grasses in China 41: 285–291.

Tripathi RS, Yadav AS, Kushwaha SPS (2012) Biology of Chromolaena odorata, *Ageratina adenophora* and *Ageratina riparia*: a review. Invasive alien plants: an ecological appraisal for the Indian subcontinent 32: 43–56. https://doi.org/10.1079/9781845939076.0043

Wang R, Wang J-F, Qiu Z-J, Meng B, Wan F-H, Wang Y-Z (2011) Multiple mechanisms underlie rapid expansion of an invasive alien plant. New Phytologist 191: 828–839. https://doi.org/10.1111/j.1469-8137.2011.03720.x

Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: An underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3: 260–267. https://doi.org/10.1111/j.2041-210X.2011.00170.x

Woodward FI (1987) Climate and plant distribution. Cambridge University Press, 188 pp.

Woodward FI, Lomas M, Kelly CK (2004) Global climate and the distribution of plant biomes. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 359: 1465–1476. https://doi.org/10.1098/rstb.2004.1525

Yates KL, Bouchet PJ, Caley MJ, Mengersen K, Randin CF, Parnell S, Fielding AH, Bamford AJ, Ban S, Barbosa AM, Dormann CF, Elith J, Embling CB, Ervin GN, Fisher R, Gould S, Graf RF, Gregr EJ, Halpin PN, Heikkinen RK, Heinänen S, Jones AR, Krishnakumar PK, Lauria V, Lozano-montes H, Mannocci L, Mellin C, Mesgaran MB, Moreno-amat E, Mormede S, Novaczek E, Oppel S, Crespo GO, Peterson AT, Rapacciuolo G, Roberts JJ, Ross RE, Scales KL, Schoeman D, Snelgrove P, Sundblad G, Thuiller W, Torres LG, Verbruggen H, Wang L, Wenger S, Whittingham MJ, Zharikov Y, Zurell D, Sequeira AMM (2018) Outstanding challenges in the transferability of ecological models. Trends in Ecology & Evolution 33: 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Supplementary material 1

Variable selection using cluster analsys based on Spearman's rank corellation and UPGMA method for agglomeration

Authors: Arunava Datta, Oliver Schweiger, Ingolf Kühn

Data type: statistical data

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/neobiota.59.36299.suppl1

Supplementary material 2

Multimodel inference table

Authors: Arunava Datta, Oliver Schweiger, Ingolf Kühn

Data type: statistical data

Explanation note: Tables depicting all the component models of the best subset (i.e. models that differed by 2 or less in AIC).

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/neobiota.59.36299.suppl2

Supplementary material 3

R codes

Authors: Arunava Datta, Oliver Schweiger, Ingolf Kühn

Data type: R code (text)

Explanation note: R codes used in the paper.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/neobiota.59.36299.suppl3